Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37350

Robustness of UCB1 and Thompson Sampling under Varying Arm Counts, Nonstationary Rewards, and Corrupted Feedback

Tianrui Huang

USC Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, USA
billyhua@usc.edu

Abstract. Upper Confidence Bound 1 (UCB1) and Thompson Sampling (TS) are two canonical algorithms for the multi-armed bandit problem. However, most empirical comparisons of their relative performance are conducted under idealized, stationary, and carefully controlled environments, leaving their robustness under real-world perturbations underexplored. This paper systematically evaluates the relative robustness of the two algorithms across four perturbation dimensions: increasing arm count, abrupt reward distribution shifts, misspecified exploration hyperparameters or priors, and unreliable observed feedback. We conduct controlled Bernoulli simulation experiments to test sensitivity to arm count, non-stationarity, and hyperparameter/prior misspecification, and complement synthetic tests with a genre-level bandit environment calibrated on the MovieLens 100K dataset. With the canonical UCB1 exploration coefficient $c = \sqrt{2}$, TS yields lower mean final cumulative regret across all tested arm counts and in the clean MovieLens environment (160.7 vs. 335.7 for UCB1). In contrast, UCB1 records slightly lower mean final dynamic regret following a single abrupt reversal of reward rankings (306.6 vs. 336.6). This performance reversal is considered instance-specific, as neither stationary baseline incorporates a forgetting mechanism for pre-change observations. Under symmetric feedback corruption, TS maintains lower mean final regret across all tested flip probabilities, but the performance gap between UCB1 and TS narrows monotonically from 174.9 at zero corruption to 37.8 at a flip probability of 0.4. Since symmetric flipping retains the Bernoulli nature of observations while compressing gaps between arm means, this narrowing is consistent with reduced signal informativeness rather than a breakdown of the Beta–Bernoulli observation model. Across the tested stationary environments, TS consistently delivers superior mean-regret performance, while non-stationary environments and unreliable feedback call for specialized extensions with drift adaptation or noise robustness.

Keywords: multi-armed bandit, UCB1, Thompson Sampling, robustness analysis, MovieLens

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA37350

1. Introduction

The multi-armed bandit (MAB) problem formalizes sequential decision-making scenarios where an agent repeatedly selects from a set of actions, observes only the reward of the chosen action, and balances exploration and exploitation to minimize cumulative regret [1]. This framework underpins a wide range of real-world systems including recommendation engines, online advertising platforms, adaptive experimentation tools, and clinical decision support systems. Upper Confidence Bound 1 (UCB1) embodies the principle of optimism in the face of uncertainty: it augments each arm's empirical mean reward with a confidence bonus, and comes with finite-time regret guarantees for bounded stochastic rewards [2]. Thompson Sampling (TS), originally proposed as a probability-matching heuristic [3], selects actions by sampling from their posterior reward distributions, and has demonstrated strong empirical performance in both simulation studies and real-world recommendation applications [4].

Nevertheless, most existing comparative studies focus on stationary environments with noise-free feedback, moderate action space sizes, and carefully tuned hyperparameters. As a result, average-case performance rankings may not generalize to realistic operational settings where environmental conditions and feedback quality vary. This paper evaluates the relative robustness of standard UCB1 and Beta–Bernoulli TS across four perturbation dimensions: growing number of arms, abrupt shifts in reward rankings, misspecified exploration coefficients or priors, and symmetric corruption of observed binary feedback. Our analysis combines controlled simulation experiments with a genre-level bandit environment constructed from the MovieLens 100K dataset.

This work makes three main contributions. First, it provides a matched synthetic benchmark where each stress factor is isolated and tested one at a time. Second, it presents a reproducible real-data experimental pipeline with explicit filtering rules and descriptive dataset statistics. Third, it disentangles reduced feedback informativeness from fundamental model misspecification when interpreting results under reward corruption. Our findings support a conditional conclusion: TS achieves lower mean final regret in the tested stationary arm-count and MovieLens environments, UCB1 yields slightly lower mean final regret in the abrupt reward-swap instance, and neither stationary baseline is designed to handle rapid environmental drift. TS also retains a mean-regret advantage under corrupted feedback, though its performance margin shrinks as observation quality degrades.

2. Related work

UCB methods select actions using optimistic estimates of uncertain rewards. UCB1 established logarithmic instance-dependent regret for bounded stochastic bandits and remains a standard baseline because it only requires empirical means and pull counts [2]. Its deterministic confidence bonus is transparent, but performance depends on the exploration scale used in practice. Modern treatments place UCB within a broader family of optimism-based stochastic and contextual bandit algorithms [1].

TS originated with Thompson's Bayesian probability-matching rule [3]. Chapelle and Li reported that TS was highly competitive with, and sometimes superior to, UCB-style alternatives on simulated, advertising, and news-recommendation data [4]. Subsequent analysis established logarithmic regret for stochastic Bernoulli bandits, and later tutorials clarified its Bayesian and frequentist interpretations [5, 6]. Recent studies continue to report empirical advantages of TS in specialized settings, including cascading recommendation bandits, while Meta-TS addresses adaptation when the prior is learned across tasks [7, 8]. These results motivate TS as a strong

DOI: 10.54254/2755-2721/2026.BA37350

baseline, but they do not provide a matched robustness comparison of standard UCB1 and Beta–Bernoulli TS under the perturbations studied here.

Non-stationary and corrupted bandits require different safeguards from stationary baselines. Variation-budget models show that dynamic environments fundamentally change achievable regret, and discounted randomized UCB and TS variants have been proposed for time-varying linear bandits [9, 10]. In corruption-robust bandits, an adversary may alter rewards before they reach the learner, specialized algorithms are designed so regret degrades with the corruption budget [11, 12]. Those models are more general than the independent symmetric flips used in this paper.

The present study therefore occupies a narrower empirical role. It does not propose a new bandit algorithm or robustness guarantee. Instead, it holds the two baseline implementations fixed, varies one designated factor at a time, and reports where the observed mean-regret advantage of TS is stable, reduced, or reversed. In the arm-count experiment, changing K also changes the separately generated arm set and its realized gap structure, so cross- K differences are not attributed to arm count alone. This design also exposes interpretation limits that can be hidden by an aggregate winner-only comparison.

3. Background and theoretical foundation

3.1. Regret in stationary and non-stationary bandits

For a stationary K -armed bandit, arm i has an unknown but fixed mean reward μ_i. At round t, the learner chooses arm a_t and observes reward r_t. If $\mu^* = \max_i \mu_i$, the realized cumulative regret is:

$$R_s(T) = \sum_{t=1}^{T} [\mu^* - r_t] \tag{1}$$

Lower cumulative regret means that the learner has sacrificed less reward while identifying and exploiting strong arms. Sublinear growth is desirable because the average regret per round then approaches zero.

In a non-stationary bandit, the mean of arm i may change with time, denoted $\mu_i(t)$. The dynamic comparator is the best arm available at each round, $\mu^*(t) = \max_i \mu_i(t)$, and dynamic realized regret is:

$$R_{dyn}(T) = \sum_{t=1}^{T} [\mu^*(t) - r_t] \tag{2}$$

This distinction matters in the abrupt-swap experiment because the identity of the best arm changes at $T/2$. A low value indicates both effective initial learning and rapid adaptation after the change.

3.2. UCB1 and thompson sampling

After pulling every arm once, UCB1 chooses:

$$a_t = \arg\max_i [\widehat{\mu}_i(t) + c\sqrt{\frac{\ln t}{n_i(t)}}] \tag{3}$$

where $\widehat{\mu}_i(t)$ is the empirical mean, $n_i(t)$ is the number of pulls of arm i, and c controls exploration. The canonical UCB1 bonus $\sqrt{2\ln(t+1)/n_i(t)}$ corresponds to $c = \sqrt{2}$ in this

DOI: 10.54254/2755-2721/2026.BA37350

parameterization. The first term favors arms with high observed reward, the square-root bonus favors arms with limited evidence. Increasing c produces broader exploration, while a small c can commit prematurely to an apparently strong arm.

For binary rewards, TS assigns each arm a $\mathrm{Beta}(\alpha_i, \beta_i)$ posterior. At each round it samples θ_i from every arm's posterior, chooses the largest sample, and updates only the chosen arm:

$$\theta_i \sim Beta(\alpha_i, \beta_i),\ a_t = \theta_i, \tag{4}$$

$$\alpha_i \leftarrow \alpha_i + \tilde{r}_t,\ \beta_i \leftarrow \beta_i + 1 - \tilde{r}_t\ (i = a_t).$$

The posterior concentrates as successes and failures accumulate. Exploration is randomized and automatically decreases for well-observed arms, but the behavior depends on the prior and on whether the observation model represents the feedback process adequately.

4. Experimental setup

4.1. Metrics

The primary metric is realized cumulative regret over M independent runs; shaded bands show ± 1 sample standard deviation (SD, denominator $M-1$). For the misspecification grid, robustness spread is defined as:

$$\text{spread} = \max_{\mathrm{h}\in\mathrm{H}} \overline{\mathrm{R}}_{\mathrm{h}}(\mathrm{T}) - \min_{\mathrm{h}\in\mathrm{H}} \overline{\mathrm{R}}_{\mathrm{h}}(\mathrm{T}) \tag{5}$$

where H is the tested hyperparameter or prior set and $\overline{R}_h(T)$ is mean final regret under setting h. A smaller spread means less within-grid variation, but the two grids differ in meaning and scale, so spread is not an algorithm-invariant robustness score. The corruption analysis reports mean final clean-utility regret, its SD, and the mean gap $\overline{R}_{\mathrm{UCB1}}(T) - \overline{R}_{\mathrm{TS}}(T)$.

4.2. Synthetic and movielens protocols

All synthetic rewards are Bernoulli. In the arm-count experiment, $K \in \{5, 20, 100\}$, $T = 2,000$, and $M = 100$. Suboptimal means are drawn once from $\mathrm{Uniform}(0.1, 0.6)$, sorted, and the best mean is set to 0.7. In the non-stationary experiment, $K = 10$, $T = 3,000$, and $M = 100$, means are evenly spaced from 0.2 to 0.6 and their ordering reverses at $T/2$. In the sensitivity experiment, $K = 10$, $T = 2,000$, and $M = 100$, UCB1 uses $c \in \{0.2, \sqrt{2}, 2.0, 8.0\}$, and TS uses Beta priors $(1, 1)$, $(1, 20)$, and $(20, 1)$. Unless deliberately perturbed, UCB1 uses the canonical coefficient $c = \sqrt{2}$ and TS uses $\mathrm{Beta}(1, 1)$.

The real-data protocol uses MovieLens 100K (100,000 ratings by 943 users on 1,682 movies) [13]. Each rating is assigned to every listed genre, binarized at four stars, and sampled with replacement from that genre's pool. A 200-event cutoff excludes only the "unknown" label (10 events) and retains all 18 named genres. Genre arms reduce sparsity and the 1,682-movie initialization burden, but they lose item-level personalization and create overlapping pools for multi-genre movies. MovieLens experiments use $T = 3,000$ and $M = 80$.

DOI: 10.54254/2755-2721/2026.BA37350

Table 1. MovieLens 100K Genre-arm Statistics after the ≥ 200 -event filter

Genre	Rating events	Success rate (rating ≥4)	Genre	Rating events	Success rate (rating ≥4)
Action	25,589	0.529	Adventure	13,753	0.534
Animation	3,605	0.563	Children	7,182	0.477
Comedy	29,832	0.501	Crime	8,055	0.596
Documentary	758	0.629	Drama	39,895	0.617
Fantasy	1,352	0.413	Film-Noir	1,733	0.706
Horror	5,317	0.468	Musical	4,954	0.540
Mystery	5,245	0.602	Romance	19,461	0.588
Sci-Fi	12,730	0.565	Thriller	21,872	0.545
War	9,398	0.665	Western	1,854	0.574

4.3. Feedback-corruption model

The corruption test separates the latent reward r_t from the observation $\tilde{r}_t$ used to update the algorithm. Independently at each pull, $\tilde{r}_t = 1 - r_t$ with probability ρ and $\tilde{r}_t = r_t$ otherwise, for $\rho \in \{0, 0.1, 0.2, 0.3, 0.4\}$. Regret is computed from the uncorrupted reward r_t, so the metric reflects lost clean utility while the learner receives unreliable feedback. This process represents accidental miscoding, mistaken ratings, or non-targeted label noise. If an arm has clean success probability p, its observed success probability is:

$$\mathrm{q} = \Pr(\tilde{\mathrm{r}}_{\mathrm{t}} = 1) = \rho + (1 - 2\rho)\mathrm{p} \tag{6}$$

For $\rho < 0.5$, ordering is preserved but every arm-mean gap is multiplied by $1 - 2\rho$. The feedback becomes less informative as ρ increases. Importantly, symmetric flipping still produces Bernoulli observations, it is not a test of a non-Bernoulli likelihood or an adaptive malicious adversary. Targeted fake reviews, arm-dependent errors, and strategic attacks require separate corruption models.

4.4. Implementation and reproducibility

Experiments are implemented in Python with NumPy and Pandas, and figures are produced with Matplotlib. Reward generation, feedback flipping, and algorithmic sampling use separate seeded random-number generators. Each Monte Carlo run receives a deterministic seed, and UCB1 and TS are evaluated under matched environment settings. MovieLens clean and corrupted experiments use the same 80 seeds (2 through 81), so the zero-corruption group is identical in both analyses. In the corruption experiment, each algorithm updates from $\tilde{r}_t$ while regret is computed from r_t. All reported parameter values are specified above.

DOI: 10.54254/2755-2721/2026.BA37350

5. Results and discussion

5.1. Robustness to the number of arms

Figure 1 compares cumulative regret for $K \in \{5, 20, 100\}$. Final mean regret for UCB1 is 107.4, 409.6, and 597.4, respectively, compared with 27.3, 116.2, and 455.8 for TS. TS therefore has lower mean final regret at every tested arm count. The margin is especially large for $K = 5$ and $K = 20$, at $K = 100$ both methods spend a larger fraction of the 2,000-round horizon collecting information. UCB1 must pull every arm once before using its index, which creates a fixed initialization cost as K grows. Because each K uses a separately generated arm set, however, the cross- K differences also reflect changes in the realized arm-gap structure and should not be attributed to arm count or initialization alone. In these tested instances with reliable binary feedback, the mean-regret result favors TS, while also suggesting that hierarchical or contextual structure would be preferable to treating a very large catalog as unrelated arms.

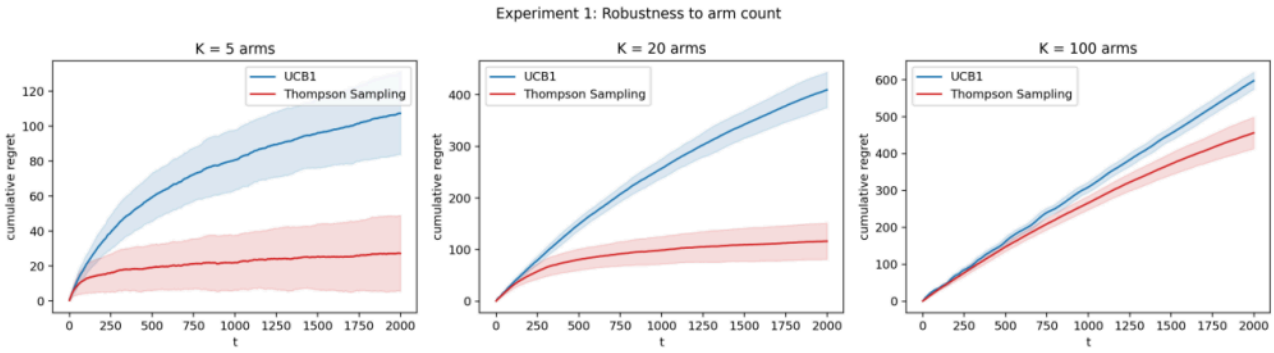

Figure 1. Cumulative regret under increasing arm count: (a) K=5, (b) K=20, and (c) K=100

Notes: Lines show means and shaded regions show ±1SD across 100 runs

5.2. Abrupt non-stationarity

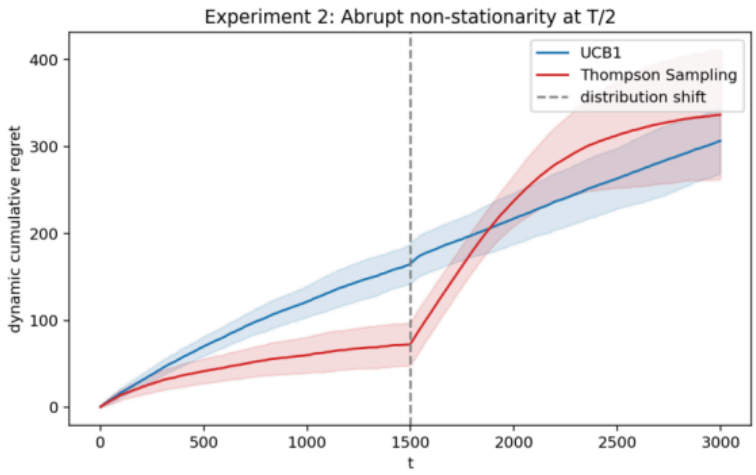

Figure 2. Dynamic cumulative regret after the arm ordering reverses at T/2

Note: Lines show means and shaded regions show ±1SD across 100 runs.

DOI: 10.54254/2755-2721/2026.BA37350

Figure 2 shows the reversal of arm ordering at $T/2$. Both algorithms accumulate a sharp additional regret after the change and then continue with approximately linear growth, final mean dynamic regret is 306.6 for UCB1 and 336.6 for TS. UCB1 has lower mean final dynamic regret in this instance, but neither baseline forgets obsolete evidence, so both continue to use beliefs learned before the swap. Because UCB behavior is sensitive to its exploration coefficient, this single ordering should not be generalized as a universal adaptation advantage. The operational implication is shared weakness: when user preferences or market conditions drift, discounted, sliding-window, restart-based, or change-detection variants are more appropriate than either stationary baseline.

5.3. Hyperparameter and prior sensitivity

Figure 3 reports the selected sensitivity grids. UCB1 mean final regret is 69.5, 237.6, 322.5, and 478.2 across $c \in \{0.2, \sqrt{2}, 2.0, 8.0\}$, giving a spread of 408.7. TS mean final regret is 70.5, 181.4, and 177.9 across priors $\mathrm{Beta}(1,1)$, $\mathrm{Beta}(1,20)$, and $\mathrm{Beta}(20,1)$, giving a spread of 110.9. Within these selected grids, TS varies less than UCB1. The spreads are descriptive rather than directly comparable robustness scores because the grids contain different numbers of settings and their parameter scales are not commensurate. The large run-to-run SD at $c = 0.2$ is consistent with occasional premature commitment, while the steadily increasing regret at $c = 8.0$ is consistent with sustained costly exploration over the tested horizon.

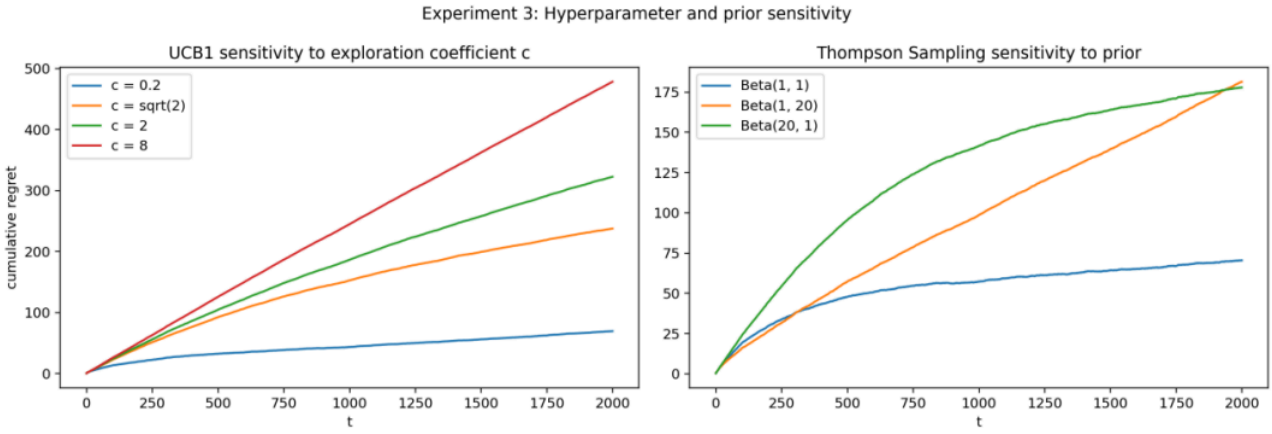

Figure 3. Sensitivity to Algorithm Settings: (a) UCB1 Exploration Coefficient c and (b) TS Beta Prior

Note: Each curve is the mean across 100 runs.

5.4. Movielens with clean feedback

Figure 4 transfers the stationary comparison to genre-level MovieLens rewards. With clean feedback, mean final cumulative regret is 335.7 for UCB1 and 160.7 for TS. The mean-regret ordering observed in the arm-count experiments therefore persists when reward probabilities are estimated from historical rating behavior rather than selected by the experimenter. Film-Noir has the highest success rate (0.706), while Fantasy has the lowest (0.413), so the environment contains meaningful but uneven arm gaps. This is a stochastic bandit calibrated from real data, not an online policy evaluation, and the conclusion remains limited to population-level genre choice rather than personalized recommendation.

DOI: 10.54254/2755-2721/2026.BA37350

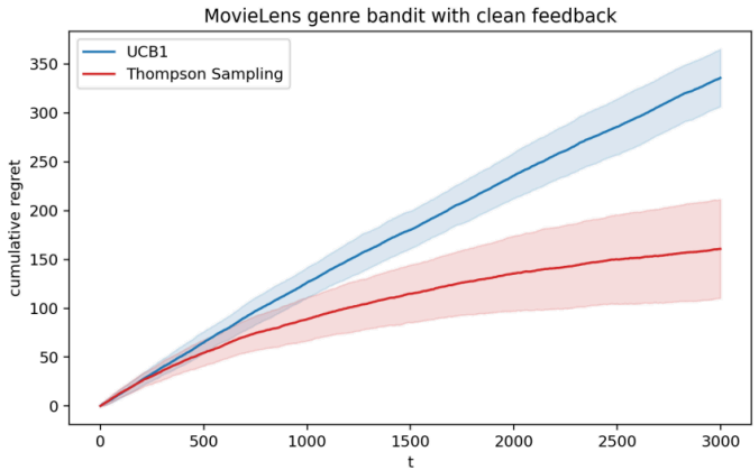

Figure 4. Cumulative regret on the movielens genre bandit with clean feedback

Note: Lines show means and shaded regions show ±1SD across 80 runs.

5.5. Movielens with corrupted feedback

Table 2 and Figure 5 present the corrected corruption experiment. Both algorithms learn from flipped observations, while regret is evaluated using the latent uncorrupted rewards. TS has lower mean final regret at every tested ρ, but its mean advantage decreases monotonically: the mean final-regret gap falls from 174.9 at $\rho = 0$ to 37.8 at $\rho = 0.4$. The zero-corruption row exactly matches the clean experiment because both analyses use the same seeds. The SDs show substantial run-to-run variability for TS, especially at moderate corruption. At $\rho = 0.4$, observed arm gaps are only 20% of their clean values, making the best genre more difficult to distinguish within the tested 3,000-round horizon.

Table 2. Final clean-utility regret under symmetric feedback corruption (mean ± SD across 80 runs)

Flip probability ρ	UCB1	Thompson Sampling	Gap (UCB1 − TS)
0.0	335.7 ± 29.3	160.7 ± 50.4	174.9
0.1	352.5 ± 29.2	227.4 ± 52.4	125.1
0.2	380.2 ± 33.2	278.6 ± 60.2	101.7
0.3	391.9 ± 27.4	325.4 ± 49.6	66.5
0.4	412.4 ± 31.4	374.6 ± 43.0	37.8

The practical interpretation is conditional. Reliable binary feedback corresponds to the largest observed mean-regret advantage for TS, while noisy feedback compresses arm differences and narrows that advantage. UCB1 does not overtake TS in mean final regret within the tested range, so the experiment does not identify a mean-regret crossover or justify selecting UCB1 solely because feedback is noisy. It instead motivates explicit noise estimation, robust updates, repeated validation of reward instrumentation, and stronger corruption models when accidental or strategic feedback errors are plausible.

DOI: 10.54254/2755-2721/2026.BA37350

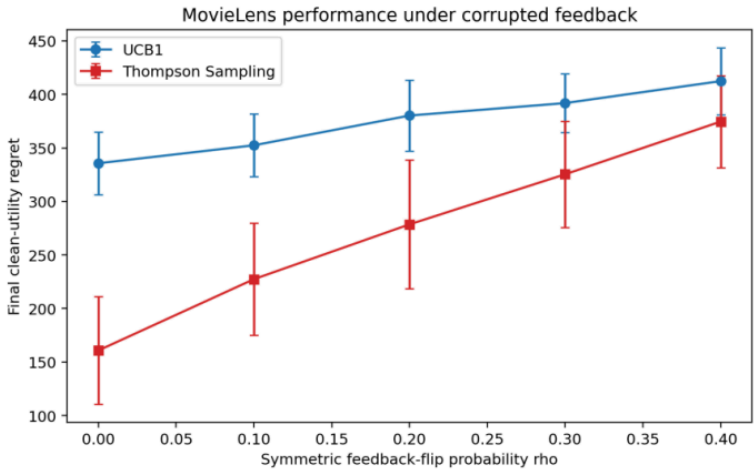

Figure 5. Final clean-utility regret under symmetric feedback flips

Note: Error bars show ±1SD across 80 runs.

6. Conclusion

This study presents a robustness-focused comparison of canonical UCB1 and Thompson Sampling, moving beyond evaluation on a single clean benchmark. TS delivers lower mean final cumulative regret across all tested arm counts and on the MovieLens 100K genre bandit, while UCB1 records lower mean final dynamic regret in the abrupt reward-swap instance. This latter finding does not indicate inherently better adaptive capability, as neither stationary baseline incorporates a mechanism to discount outdated observations, and UCB performance is sensitive to its exploration coefficient setting. Across the selected sensitivity grids, TS exhibits smaller performance variation than UCB1, though the non-commensurate parameter grids preclude a direct universal robustness comparison. Under symmetric feedback corruption, TS maintains lower mean final regret across the entire tested range, but its performance margin over UCB1 narrows from 174.9 to 37.8 as the flip probability rises from 0 to 0.4. This convergence pattern aligns with reduced signal informativeness: symmetric flipping preserves the Bernoulli structure of observations while compressing gaps between arm means, and thus the results do not indicate a fundamental failure of the Beta–Bernoulli observation model. Our conclusions are bounded by the scope of the experimental design. The real-data analysis is based on a single historical dataset, uses population-level genre arms with overlapping membership, applies a binary rating threshold, operates over relatively short horizons, and assumes independent symmetric noise. We compare only standard UCB1 and Beta–Bernoulli TS, and the sensitivity analysis relies on small, non-commensurate parameter grids. While these design choices enable a transparent baseline comparison, they limit external validity and do not capture settings involving personalization, contextual recommendation, targeted adversarial manipulation, delayed feedback, heavy-tailed rewards, or gradual environmental drift.

Future research can be extended in several directions. First, evaluation should be expanded to multiple recommendation and advertising datasets, and contextual features can be incorporated to model item-level or user-level heterogeneity. Second, in addition to mean regret and standard deviation, researchers should report normalized regret and high-probability regret bounds for more comprehensive performance characterization. Third, more realistic corruption patterns should be tested, including arm-dependent noise, bursty errors, and adaptive adversarial attacks. Finally, drift-aware variants of UCB and TS, robust posterior update mechanisms, and corruption-robust bandit

DOI: 10.54254/2755-2721/2026.BA37350

algorithms should be compared under identical random seeds and consistent utility definitions for fair benchmarking. Within the scope of this work, mean-regret results consistently favor TS in the tested stationary environments with reliable binary feedback. For dynamic environments and unreliable observation settings, specialized algorithms with explicit forgetting mechanisms, noise resilience, or adversarial robustness are more appropriate.

References

[1] Lattimore, T., & Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press. doi: 10.1017/9781108571401.

[2] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning, 47*(2–3), 235–256. doi: 10.1023/A:1013689704352.

[3] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika, 25*(3–4), 285–294. doi: 10.1093/biomet/25.3-4.285.

[4] Chapelle, O., & Li, L. (2011). An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*.

[5] Agrawal, S., & Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory, PMLR, 23*, 39.1–39.26.

[6] Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning, 11*(1), 1–96. https://doi.org/10.1561/2200000070

[7] Zhong, Z., Cheung, W. C., & Tan, V. Y. F. (2021). Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research, 22*(218), 1–66.

[8] Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-W., Mladenov, M., Boutilier, C., & Szepesvári, C. (2021). Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning, PMLR, 139*, 5884–5893.

[9] Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. In *Advances in Neural Information Processing Systems 27*.

[10] Kim, B., & Tewari, A. (2020). Randomized exploration for non-stationary stochastic linear bandits. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, PMLR, 124*, 71–80.

[11] Lykouris, T., Mirrokni, V., & Paes Leme, R. (2018). Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* (pp. 114–122).

[12] Gupta, A., Koren, T., & Talwar, K. (2019). Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the Thirty-Second Conference on Learning Theory, PMLR, 99* (pp. 1562–1578).

[13] Harper, F. M., & Konstan, J. A. (2015). The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems, 5*(4), Art. 19, 19:1–19:19. https://doi.org/10.1145/2827872